In previous lecture, we proved that using \( N = O(n/\epsilon^2) \) samples from an unstructured distribution \( P \) over \( n \) elements, we can output a distribution \( Q \) such that with constant probability, the total variational distance between \( P \) and \( Q \) is small i.e. \( d_{TV}(P,Q) \leq \epsilon \). In this lecture, we prove a matching lower bound showing that \( \Omega(n/\epsilon^2) \) samples are necessary for such an algorithm to succeed with constant probability.

### 2.1 Notation

\( H(P), H(PQ), H(P|Q) \) and \( I(P;Q) \) are the entropy, joint entropy, conditional entropy and mutual information of random variables \( P \) and \( Q \) respectively (see previous lecture for definitions).

### 2.2 Standard Theorems from Information Theory

Apart from the theorems discussed last lecture, we will also use the following standard theorems.

**Theorem 2.1 (Shanon Coding Theorem (Informal))**  To encode \( n \) i.i.d. copies of random variable \( P \), we need \( \approx n \cdot H(P) \) bits.

Informally, the following theorem says that for two random variables \( P \) and \( Q \), we can not increase their mutual information by running a deterministic function over them.

**Theorem 2.2 (Information Processing Inequality)**  For any deterministic function \( f \),

\[
I(P;Q) \geq I(f(P);Q) \tag{2.1}
\]

**Proof:**

\[
I(P;Q) = I(Pf(P);Q) \tag{2.2}
\]
\[
= I(f(P);Q) + I(P;Q|f(P)) \tag{2.3}
\]
\[
\geq I(f(P);Q) \tag{2.4}
\]

**Theorem 2.3**  For random variables \( Y \) and \( X_1, \ldots, X_n \) such that \( X_i \) is independent of \( X_1, \ldots, X_{i-1} \) given \( Y \), we have that

\[
I(Y;X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} I(Y;X_i) \tag{2.5}
\]

**Proof:**  For random variable \( A, B, C \) where \( A \) and \( B \) are independent given \( C \) i.e.

\[
I(A;B|C) = 0 \tag{2.6}
\]
we have
\[ I(A; C) = I(A; BC) - I(A; B|C) = I(A; B) + I(A; C|B) \geq I(A; B) \quad (2.7) \]

Using this, we get
\[
I(Y; X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} I(Y; X_i | X_1, \ldots, X_{i-1}) \leq \sum_{i=1}^{n} I(Y; X_i) \quad (2.8)
\]

\[ \leq \sum_{i=1}^{n} I(Y; X_i) \quad (2.9) \]

### 2.3 Lower Bound

We first formally state the lower bound:

**Theorem 2.4** There exists a unstructured distribution \( P \) over \( 2n \) elements such that any algorithm which returns a distribution \( Q \) such that \( d_{TV}(P, Q) < \epsilon \) requires \( \Omega(n/\epsilon^2) \) samples from \( P \).

**Proof:** Using Yao’s minimax principle, it is sufficient to show that there exists an ensemble of distributions \( D \) s.t. an adversary can choose a specific distribution \( P \in D \) and any algorithm (with knowledge of \( D \) but unaware of \( P \)) requires at least \( O(n/\epsilon^2) \) samples to output a distribution \( Q \) such that \( d_{TV}(P, Q) < \epsilon \) with constant probability.

For our ensemble of distributions, we consider \( n \) random variables \( \{X_i\}_{i=1}^{n} \), each \( X_i \in \{-1, 1\} \) and corresponding pairs of bins where the probabilities for the \( i^{th} \) bin are given by
\[
\left( \frac{1 + 6X_i \epsilon}{2n}, \frac{1 - 6X_i \epsilon}{2n} \right) \quad (2.10)
\]

Note that the probabilities are normalized.

Let \( Y_i \) be the estimate of \( X_i \). Each \( i \) for which \( X_i \neq Y_i \) contributes at least \( 6\epsilon/2n \) to \( d_{TV}(P, Q) \). Therefore, if we have \( Q \) s.t. \( d_{TV}(P, Q) \leq \epsilon \), then \( X_i = Y_i \) on at least \( 2/3 \) of the coordinates.

Let
\[
C = \begin{cases} 
1 & d_{TV}(P, Q) < \epsilon \\
0 & \text{otherwise} \end{cases} \quad (2.11)
\]

We first recall that the mutual information between random variables \( X \) and \( Y \) is defined as \( H(X) - H(X|Y) \) and is denoted as \( I(X; Y) \).

The main idea is that when the algorithm succeeds (i.e. \( C = 1 \)), the value of \( Y \) gives us some information about \( X \). In particular, if the algorithm succeeds with constant probability (or we can reliably learn \( X \)) then \( I(X; Y) = \Omega(n) \). Also, we will show that \( I(X; Y) = O(N\epsilon^2) \), which would imply the lower bound.

More formally, we know
\[
I(X; Y) = I(X; C) + I(X; Y|C) \quad (2.12)
\]

\[ \implies I(X; Y) \geq I(X; Y) - 1 \quad (2.13) \]

\[ = H(X) - H(X|Y) - 1 \quad (2.14) \]
where the second last step follows from $I(X; C) \leq 1$ and $I(X; Y) \geq I(X; Y|C)$.

From Theorem 2.1, we get

$$H(X) \leq n \tag{2.15}$$

Also, we can write $H(X|YC)$ as

$$H(X|YC) = \Pr[C = 0]H((X|C = 0)|(Y|C = 0)) + \Pr[C = 1]H((X|C = 1)|(Y|C = 1)) \leq \Pr[C = 0]n + \Pr[C = 1]0.9n \tag{2.16}$$

Where the last line is because if $C = 1$ then $X$ and $Y$ agree on $2n/3$ bits. Hence the conditional entropy of $X$ is at most $\log \left( \sum_{i=0}^{n/3} \binom{n}{i} \right)$.

Also, by Theorem 2.3,

$$I(X; S_1, \ldots, S_N) \leq \sum_{i=1}^{n} I(X; S_i) = N \cdot I(X; S) \tag{2.18}$$

We can think of this as $X$ picks $n$ biases for coins, $S$ picks a random coin $A$ and flips it to return $t$. Hence, we can write $I(X; S)$ as

$$I(X; S) = I(X; At) = I(X; A) + I(X; t|A) \leq \theta(\epsilon^2) \tag{2.19}$$

$$= I(X_A; t) \leq \theta(\epsilon^2) \tag{2.20}$$

where the last step follows from previous lecture. Substituting Equation (2.15), (2.17), (2.18) and (2.21) in Equation (2.14) gives (also using that $Y$ is a deterministic function of $S_1; \ldots, S_n$)

$$N\theta(\epsilon^2) \geq NI(X; S) \geq \Pr[C = 1] \frac{n}{10} - 1 \tag{2.22}$$

Therefore, for constant probability of success, we need $N = \Omega(n/\epsilon^2)$. ■